Lightbeam

Lightbeam Data Identity Graph

Rewriting the Rules of Data Security

WHITEPAPER

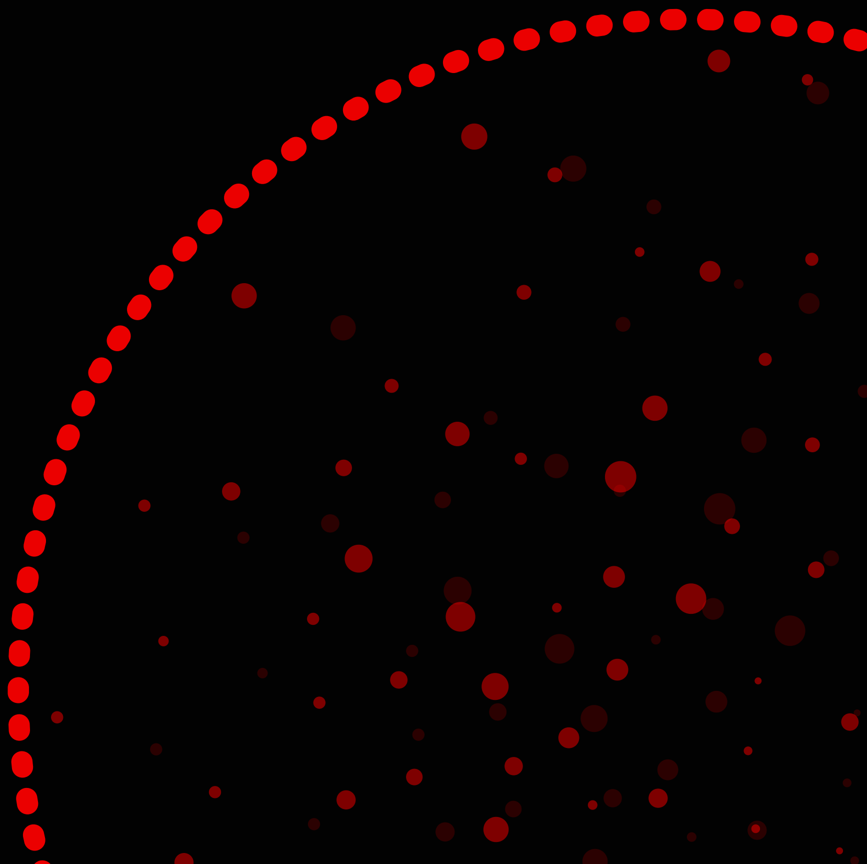


Table of Contents

Executive Summary 03

Introduction 04

Defining the Data Identity Graph 05

The Architecture Behind the Data Identity Graph 08

Scalability and Performance at Petabyte Scale 12

The Connections that Power Governance 15

Governance and Policy Enforcement 16

Enhancing Privacy Operations with the Data Identity Graph 17

Real-World Impact of the Data Identity Graph 19

How The Lightbeam Data Identity Graph Is Different 22

Future Direction and Innovation 23

Conclusion 26

Glossary of Terms 27

About Lightbeam 30

Executive Summary

The largest risks and costs in cybersecurity stem from the exposure of sensitive data, whether through data breaches, ransomware, or violations of privacy regulations. Yet despite heavy investments in perimeter security and compliance tools, businesses continue to struggle with data security. The core problem is simple: sensitive data elements are often dynamic, stored in disparate locations, and frequently in unstructured formats. Sensitive content is created, copied, and shared by employees whose priority is getting work done, not adhering to governance policies.

Without real-time understanding of what data is sensitive, whose data it is, where it's located, and who has access to it, organizations can't effectively secure it. Written policies alone are insufficient when most teams can't even identify which files contain sensitive data, much less understand the context of ownership or business function. The solution isn't more alerts or dashboards. What's needed is a continuously updated system, accurate enough to autonomously change access policies, redact sensitive information, and trigger incident response without human intervention.

That system is the Lightbeam Data Identity Graph. Built on patented entity resolution and context-aware AI, the Data Identity Graph redefines data security, governance, and privacy by linking sensitive data to the people, assets, and business processes it represents. This identity-centric foundation enables not only automated governance, but also automated remediation, allowing access to be revoked, data to be quarantined or redacted, and incidents to be contained in real time. As a result, CISOs and security teams gain continuous visibility, precise control, and the ability to reduce risk autonomously across customer, employee, and business-critical data.

Introduction: The Identity Crisis in Data Security

The stakes in data security have never been higher. Every high-profile breach, ransomware attack, and regulatory fine ultimately traces back to the same root cause: sensitive data was exposed. It was accessible when it shouldn't have been. It wasn't protected because no one knew it was there or who it belonged to.

In today's enterprise, sensitive data exists everywhere. It lives in emails, spreadsheets, chat threads, PDFs, and AI-generated documents. It's created and shared by employees focused on productivity, not policy. It moves across systems, is duplicated for convenience, and often stored indefinitely. It's a living, breathing part of business operations, but most security and governance frameworks treat it like a static asset.

Worse, organizations are being asked to enforce access controls and privacy regulations without knowing the basic facts: What files contain sensitive data? Whose data is it? How does that person relate to the business? What permissions are appropriate, and which ones are excessive?

Traditional tools weren't built for this level of context. They scan for patterns, flag files, and issue alerts, but they don't understand the context behind the data. And without that understanding, it's impossible to secure sensitive information at scale.

To solve this, businesses need more than classification. They need comprehension. They need a system that builds a continuously updated, identity-centric view of their data estate, one that is accurate enough to automate access policy enforcement, redact at-risk content, and trigger incident response without manual review. That is why Lightbeam built the Data Identity Graph.

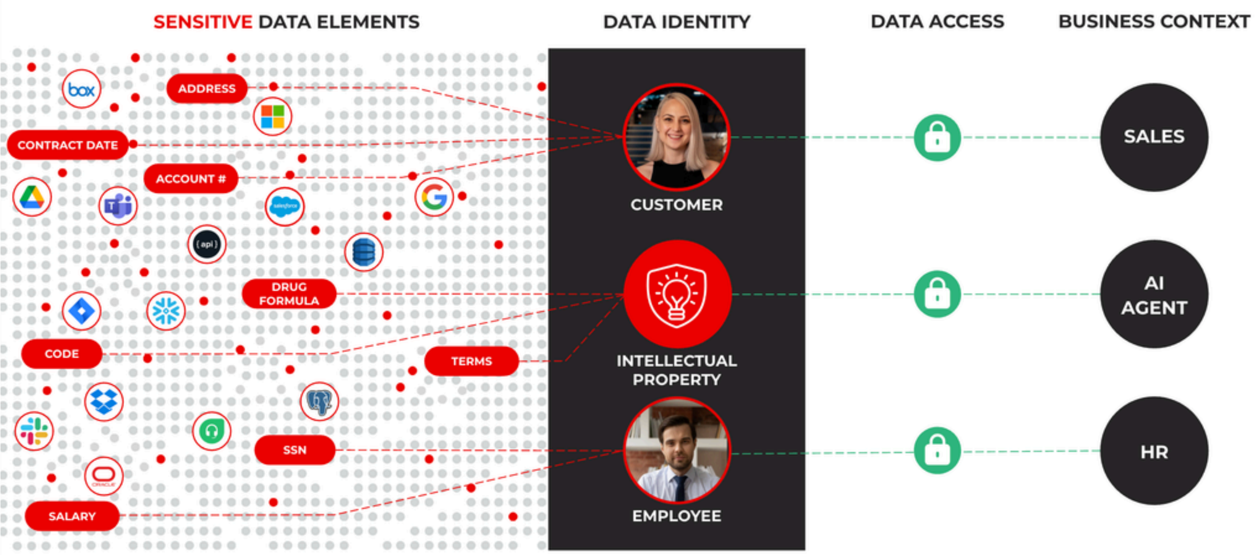
At the core of the Data Identity Graph is the concept of an entity.
An entity represents a data subject, which is any person, group, or object
that sensitive data is about. While many entities are human identities such as
employees, customers, or patients, they can also be non-human subjects like
companies, departments, contracts, applications, or even digital assets such as files or service accounts. The term entity reflects this broader definition, allowing the system to associate sensitive data with the full range of data identities that matter in security, privacy, and governance.

By modeling data relationships around data identities rather than just documents or fields, the Data Identity Graph enables richer, more accurate context, connecting data to the people and organizations it's associated with as well as the systems and assets it touches. This entity-aware architecture is what allows Lightbeam to deliver identity-driven data security at scale.

The Lightbeam Data Identity Graph doesn't just show what's in a file, AI chat, Salesforce.com record, or email. It reveals whose data it is, who can access it, and whether that access is appropriate based on role, behavior, and policy. It becomes the foundation for a new kind of data security, one that's precise, proactive, and reduces risk continuously and autonomously.

Defining the Data Identity Graph

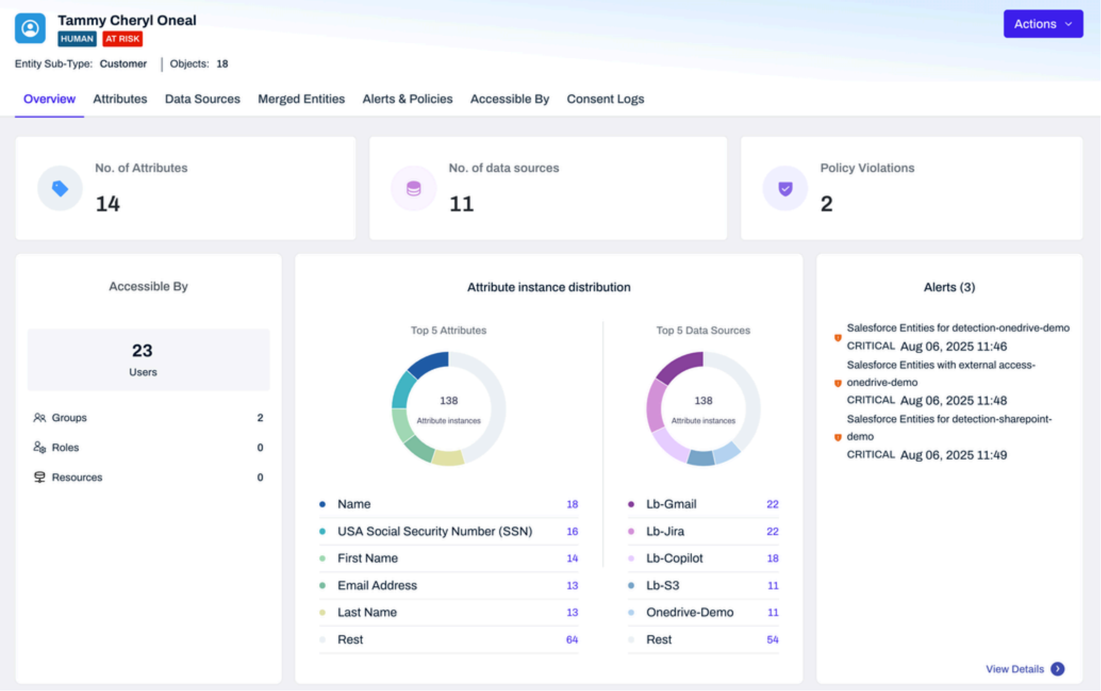
The Data Identity Graph is Lightbeam's core intelligence layer. It is a dynamic, AI-powered system that continuously links fragmented sensitive data elements such as names, emails, account numbers, and medical record numbers to entities like customers, partners, contracts, or intellectual property. As it processes data from structured systems (such as databases), semi-structured sources (like CSVs and logs), and unstructured content (including documents, chats, and emails), it builds a holistic and continuously evolving model of how disparate data points connect to data entities, enriched with full business context.



Depiction of the connections and relationships the Data Identity Graph draws between discovered sensitive data elements, the data identity, and the data access paths that lead to those data identities.

Rather than operate as a static map, the Data Identity Graph continuously learns and refines its understanding by integrating new data elements and applying patented entity resolution methods. As it scans more sources over time, the system incrementally improves the accuracy and richness of each data entity profile. For example, an initial scan might detect a name and email address linked to a financial advisor in a CRM record. Later scans of HR rosters, file shares, and email threads might reveal the same individual's employee ID, phone number, department affiliation, and access history. The graph merges these data elements into a single, unified entity. This enables more precise classification, access policy enforcement, and anomaly detection as the model evolves. This continuous enrichment allows the system to resolve data identities even when incoming data is partial, noisy, or duplicated across environments, dramatically improving data classification accuracy to over 96 percent.

The graph models two primary roles for entities: data subject entities and accessor entities. Data subject entities represent the subject of the data, whether that is a person, a company, a contract, a department, or any other digital asset that the data describes or is associated with. Accessor entities refer to the users, service accounts, or systems that interact (or could interact) with that data. These relationships are further enriched with business context such as document location, department ownership, access patterns, timestamps, and source system. This allows Lightbeam to understand not just whose data it is or what the data is about, but also how it is being used and accessed.



Tammy Cheryl Oneal
HUMAN AT RISK
Entity Sub-Type: Customer | Objects: 18

No. of Attributes 14

No. of data sources 11

Policy Violations 2

Accessible By

23 Users

Groups	2
Roles	0
Resources	0

Attribute instance distribution

Top 5 Attributes — 138 Attribute instances

Name	18
USA Social Security Number (SSN)	16
First Name	14
Email Address	13
Last Name	13
Rest	64

Top 5 Data Sources — 138 Attribute instances

Lb-Gmail	22
Lb-Jira	22
Lb-Copilot	18
Lb-S3	11
Onedrive-Demo	11
Rest	54

Alerts (3)

Salesforce Entities for detection-onedrive-demo
CRITICAL Aug 06, 2025 11:46

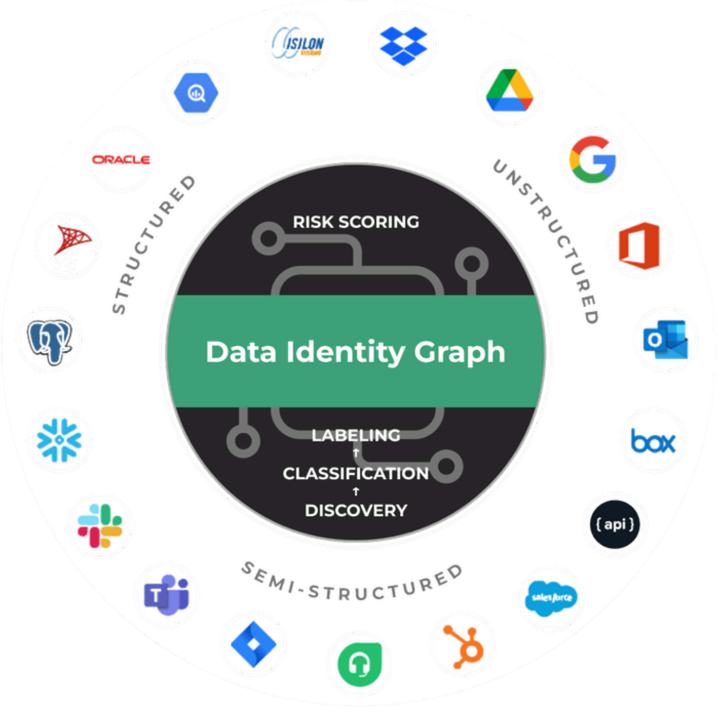
Salesforce Entities with external access-onedrive-demo
CRITICAL Aug 06, 2025 11:48

Salesforce Entities for detection-sharepoint-demo
CRITICAL Aug 06, 2025 11:49

As a result, the Data Identity Graph provides not just visibility, but actionable insight. It understands that a six-figure number in a spreadsheet is not simply a financial value. It recognizes that it represents the annual salary of Jane Smith, a senior engineer. Her compensation appears in a budget planning file that was inadvertently copied from a restricted HR system into a shared project folder. That folder, in turn, is accessible to the entire project team, not just her manager and HR. This is not just a case of misclassified data. It is a privacy and security incident with real consequences for both the employee and the organization.

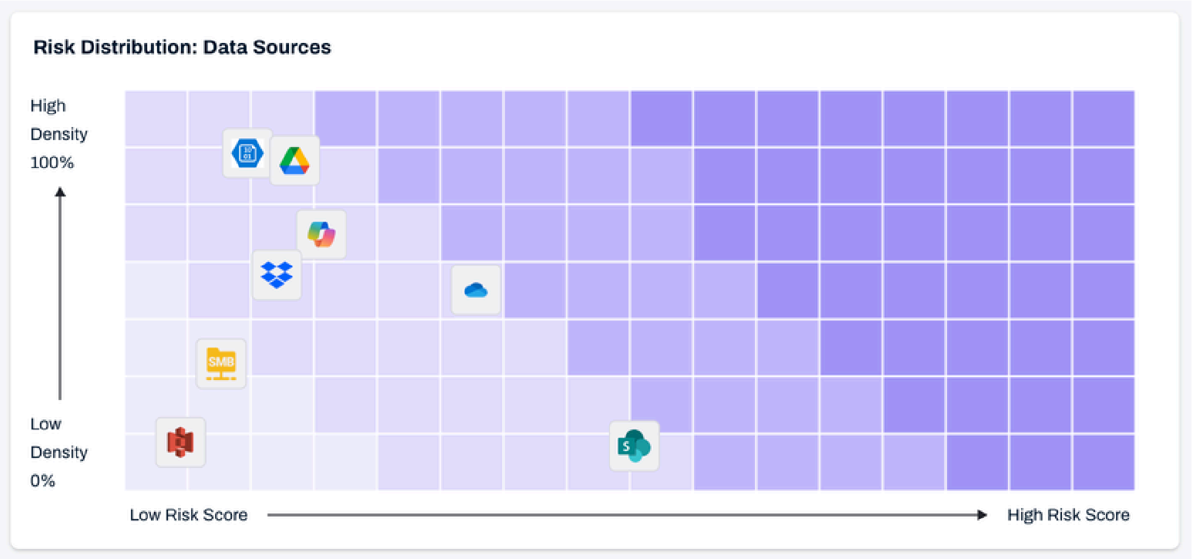
Data Inputs and Monitoring Scope

The Lightbeam Platform's ability to ingest and analyze data spans the entire digital footprint of the enterprise. It connects to data sources regardless of format or structure. Where a defined schema exists—such as in SQL databases, HR platforms, or CRMs—Lightbeam utilizes it to enhance context and precision. However, the platform does not require a schema to function. It is equally effective at analyzing semi-structured business data (e.g., Jira tickets, JSON logs) and completely unstructured content, including PDFs, Word files, Slack conversations, and Microsoft Copilot sessions.



Circle showing the core foundation of the Lightbeam platform: Data Discovery, Data Classification, Data Labeling, Risk Scoring, and the Data Identity Graph.

Each data element is parsed for identifiable attributes. These attributes are then correlated to known data subject identities within the graph. Unlike traditional scanning engines that only match patterns, Lightbeam's AI models evaluate context, sensitivity, and business relationships to determine whether the attribute is accurate, current, and meaningful. This correlation improves over time, as the graph accumulates more data points, reinforces patterns, and refines ambiguous or conflicting matches.



Screenshot of the Risk Distribution of Data Sources, showing multiple data sources ranked on two axis from Low Density to High Density on Y-axis, Low Risk Score to High Risk Score on X-axis.

Based on the sensitive data elements mapped in the Data Identity Graph, Lightbeam prioritizes risk by evaluating both the volume of sensitive data within a datastore and the density of risk concentrated inside it. Rather than treating all repositories equally, the platform identifies where sensitive data is most heavily clustered and where identity and access patterns increase the likelihood of exposure. A smaller datastore with high-risk density can be prioritized over a larger system with limited sensitive content. This identity-centric approach enables security teams to focus remediation on the data stores that pose the greatest real-world risk first.

The Architecture Behind the Data Identity Graph

The Lightbeam Data Identity Graph is more than a static repository of relationships. It is an intelligent, continuously evolving system built to map sensitive data to data identities with real-world business context at speed and scale. At its core, the architecture integrates patented entity resolution, AI-powered classification, and Retrieval-Augmented Generation (RAG) to enable real-time, identity-aware data security.

The architecture begins with automated discovery and classification. Lightbeam deploys a distributed data plane to scan across structured sources such as SQL databases and CRMs, semi-structured systems like Salesforce and Jira, and unstructured content including PDFs, emails, chat logs, images, and documents. As data is ingested, it is parsed for sensitive data elements such as names, account numbers, national IDs, or medical record numbers. AI models, which have been custom trained by Lightbeam, go beyond pattern matching by evaluating the surrounding context to classify document types such as legal contracts, onboarding forms, or patient discharge summaries.

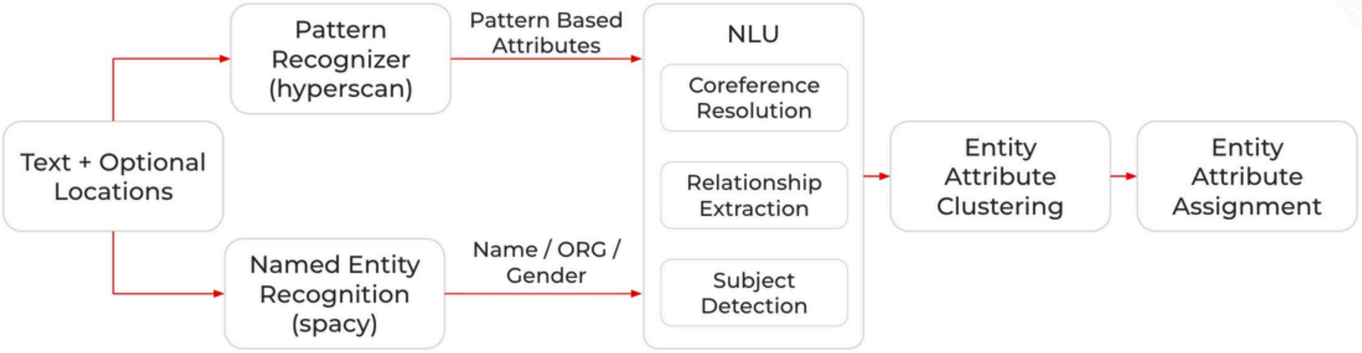
Unlike SaaS-only solutions that require data to be moved or mirrored into a central environment, Lightbeam separates the data plane from the control plane. This gives enterprises the flexibility to deploy any number of data plane scanning components as close as possible to where the data resides, across multiple cloud providers and on-premises environments. For example, data in AWS S3 can be scanned within the customer's AWS VPC, while files in Azure Blob Storage or an on-prem NAS can be scanned independently in those environments without ever leaving the network. This model minimizes bandwidth usage, preserves data sovereignty, and supports regional compliance requirements.

The control plane, which handles centralized policy configuration, reporting, and analytics, can be located wherever it is most secure and cost-effective, whether in the cloud or in a private data center. It orchestrates the scanning process and aggregates insights across all connected data planes without directly handling the underlying data.

The architecture of the Lightbeam platform also enables independent, elastic scaling of resources in each data plane location. During initial deployment or onboarding of a new datasource, resources can be scaled up to accelerate the scanning of the organization's entire historical data estate. Once this comprehensive scan of existing data is complete, the system transitions to efficient, incremental scanning that only processes newly created or modified data. At that stage, resources can be scaled down, maintaining real-time performance while significantly optimizing compute costs.

This architectural flexibility allows Lightbeam to adapt to the size, geography, and complexity of any data estate while delivering continuous discovery, classification, and governance without compromising control, performance, or cost-efficiency.

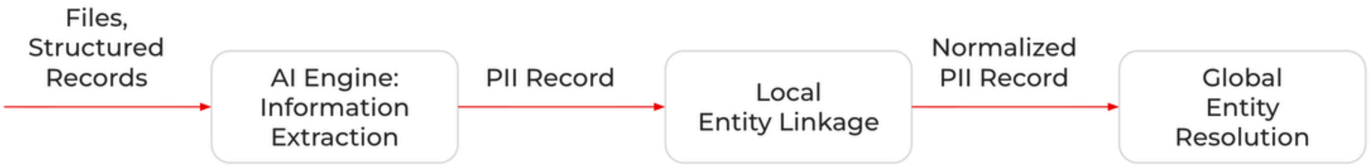
Once data is parsed, the system performs intelligent grouping. Rather than treating each identifier in isolation, it clusters related information within a document. For example, it recognizes that a name, address, and account number mentioned multiple times likely belong to the same individual. Natural language processing models track pronouns, references, and relationships across the content to form “local entities,” which are temporary identity constructs built from document context.


Text + Optional Locations
Pattern Recognizer (hyperscan)
Pattern Based Attributes
Named Entity Recognition (spacy)
Name / ORG / Gender
NLU
Coreference Resolution
Relationship Extraction
Subject Detection
Entity Attribute Clustering
Entity Attribute Assignment

Depiction of the Entity identification process through Pattern Recognition, Natural Language Understanding (NLU), Name Entity Recognition, Attribute Clustering, and more.

These local entities are then evaluated through the Lightbeam identity resolution engine. A confidence score is calculated using weighted matching across available identifiers. Definitive data elements such as Social Security numbers or Employee IDs contribute significantly to the score.

Matches against trusted Systems of Truth (SOTs), which are authoritative sources like HR systems, identity governance platforms, or customer relationship management databases, receive additional weighting because they represent the organization’s gold source for verified identity data. Systems of Truth also enable the platform to determine the subtype of the entity (e.g. For the Person type, the sub-type might be: Employee, Vendor, Partner, Customer). This process is fully automated and dynamically tuned. Administrators do not need to manually configure matching thresholds or scoring logic. If the system determines that confidence is high, the local entity is merged into an existing profile within the graph. If ambiguity remains, the system preserves it as a provisional identity, allowing for refinement as more data is ingested over time.



With the Lightbeam architecture, Retrieval-Augmented Generation (RAG) enhances the system's ability to understand the relationships between data, entities, and business context. This approach allows the platform to retrieve relevant, pre-learned templates and apply them to new, structurally similar documents or data sources, enabling accurate and context-aware classification, extraction, and entity mapping.

For example, a procurement analyst can upload a custom "New Vendor Onboarding Form" and annotate where sensitive fields (like bank account numbers or tax IDs) appear. Lightbeam stores that annotated template in its knowledge base. Later, when Lightbeam encounters a new document with a similar structure, it retrieves the template, classifies the document as a vendor agreement, and applies the same field-level logic to extract and label sensitive data with precision. And it doesn't stop at finding attributes: Lightbeam can also resolve extracted values to entities in the Data Identity Graph, grouping vendor contact details into a single "vendor" person entity (with the right subtype), so downstream policies can reason about who the data belongs to, not just what was found.

This not only enables consistent classification across variants of the same form type, but also supports tailored remediation actions such as quarantining the file or restricting access to procurement staff.

Beyond classification, the system also maps relationships between extracted data elements and specific entities. For example, a document may include information about both a new hire and a vendor. Lightbeam can distinguish between these entities, associate the correct data to each one, and assign contextual tags such as "New Hire" or "Partner." Understanding these relationships is essential for enforcing accurate governance. It allows the system to determine not only what data is sensitive, but also who or what it is about, how entities relate to each other, and how that context impacts access, retention, or remediation policies.

To achieve over 96 percent classification precision, Lightbeam uses a multi-stage process designed to balance recall and accuracy. The first phase emphasizes high recall, casting a wide net using pattern matching and lightweight machine learning models to detect any possible instance of sensitive data. This ensures that no potentially relevant information is missed during the initial scan.

In the second phase, contextual filters refine the results to reduce false positives and improve precision. This includes natural language understanding (NLU), semantic similarity models, and cross-checking against verified Systems of Truth (SOTs) such as HR systems or CRM platforms. These mechanisms validate that detected data is not only structurally correct but also contextually appropriate.

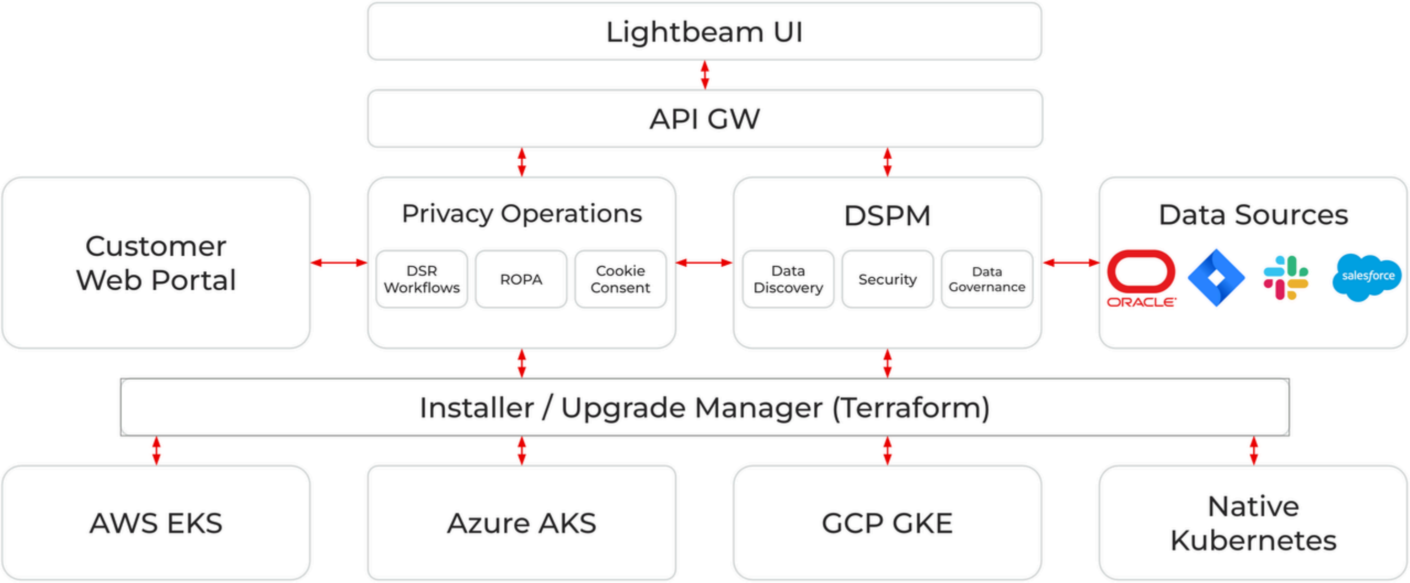
When particularly complex or ambiguous documents are encountered, Lightbeam selectively uses Large Language Models (LLMs) to provide deeper reasoning and interpretation. These LLMs are deployed locally within the customer's environment, ensuring that sensitive data never leaves the network. This approach maintains strong data privacy while minimizing latency and compute costs. By using LLMs only when necessary, Lightbeam preserves high performance without compromising the precision and depth of analysis.

This architecture is also optimized for performance and resource efficiency. Advanced filtering techniques like blocking dramatically reduce the complexity of comparisons during entity resolution. Purpose-built indexing structures allow rapid lookups across billions of records. While Lightbeam supports the use of GPUs, the models are optimized to run on standard infrastructure, with CPU-efficient transformers and distillation techniques for reduced memory footprint.

Together, these components form a real-time intelligence layer that enables automated access control, incident response, retention enforcement, and privacy workflows, all grounded in identity. The Lightbeam architecture does not simply scan data. It understands it. And by understanding, it secures.

Scalability and Performance at Petabyte Scale

At enterprise scale, data security is not a one-time inventory project. It is a continuous, context-aware process that must adapt to the speed and complexity of modern business environments. The Lightbeam Data Identity Graph is engineered specifically to meet the needs of petabyte-scale environments, with a highly optimized, multi-layered architecture designed for performance, efficiency, and precision.



High-level architectural diagram of the Lightbeam platform, showing the underlying infrastructure layer, all the way up through data sources, data discovery, privacy operations, to the Lightbeam User Interface.

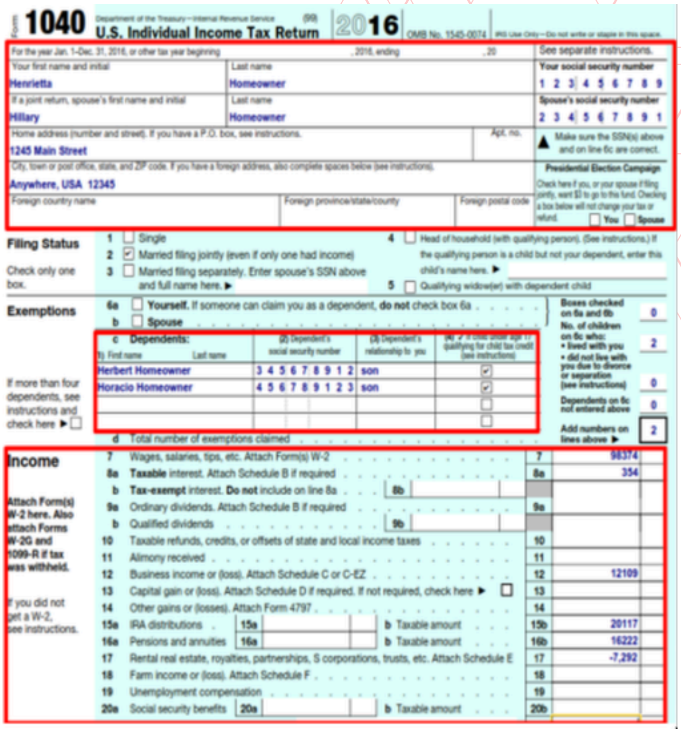
The system uses a fully distributed, scale-out architecture to ingest massive volumes of data in parallel. Multiple processing nodes are deployed directly into the environments where the data resides, whether in AWS, Azure, GCP, OCI, or on-prem systems. This ensures scanning happens close to the source, reducing network costs and preserving data sovereignty. As the size of the data estate grows, Lightbeam automatically scales horizontally by adding additional processors, eliminating bottlenecks and increasing throughput in proportion to demand.

After initial ingestion, the system transitions to incremental scanning. By listening to real-time event notifications such as S3 object changes or database updates, it processes only what has changed. This avoids unnecessary full re-scans, dramatically improving performance, reaction time to changes, and reducing cost without compromising visibility.

The Lightbeam AI pipeline is designed for scale. It's built from specialized, modular components for tasks like text detection and recognition, layout analysis, key-value pair extraction, language detection, natural language understanding, attribute extraction, and local entity clustering. Each component runs independently and can scale up or down based on where the bottleneck is. For example, if the system hits a surge in image-based documents, only the image-processing components scale out, without dragging the rest of the pipeline along. This micro-component design helps Lightbeam maximize resource efficiency while maintaining consistent performance.

Smart algorithmic filtering further accelerates entity resolution. Rather than perform computationally expensive one-to-one comparisons across all records, Lightbeam uses an approach called blocking. This technique clusters candidate matches before comparison, reducing operations by orders of magnitude while maintaining high precision.

At the core of the platform is a high-performance indexing engine built specifically for entity resolution. A custom database schema and indexing strategy enable sub-second lookups across billions of records. When evaluating multiple data elements for a match, the system returns results in milliseconds.



Form **1040** Department of the Treasury—Internal Revenue Service (99)
U.S. Individual Income Tax Return **2016** OMB No. 1545-0074 IRS Use Only—Do not write or staple in this space.

For the year Jan. 1–Dec. 31, 2016, or other tax year beginning , 2016, ending , 20 See separate instructions.

Your first name and initial Henrietta	Last name Homeowner	Your social security number 1 2 3 4 5 6 7 8 9
If a joint return, spouse's first name and initial Hillary	Last name Homeowner	Spouse's social security number 2 3 4 5 6 7 8 9 1
Home address (number and street). If you have a P.O. box, see instructions. 1245 Main Street	Apt. no.	▲ Make sure the SSN(s) above and on line 6c are correct.
City, town or post office, state, and ZIP code. If you have a foreign address, also complete spaces below (see instructions). Anywhere, USA 12345		**Presidential Election Campaign** Check here if you, or your spouse if filing jointly, want $3 to go to this fund. Checking a box below will not change your tax or refund. □ You □ Spouse
Foreign country name	Foreign province/state/county	Foreign postal code

Filing Status

Check only one box.

1 □ Single
2 ☑ Married filing jointly (even if only one had income)
3 □ Married filing separately. Enter spouse's SSN above and full name here. ▶
4 □ Head of household (with qualifying person). (See instructions.) If the qualifying person is a child but not your dependent, enter this child's name here. ▶
5 □ Qualifying widow(er) with dependent child

Exemptions

6a □ **Yourself.** If someone can claim you as a dependent, **do not** check box 6a . . . . . .
b □ **Spouse** . . . .

Boxes checked on 6a and 6b 0

c **Dependents:**

(1) First name Last name	(2) Dependent's social security number	(3) Dependent's relationship to you	(4) ✓ if child under age 17 qualifying for child tax credit (see instructions)
Herbert Homeowner	3 4 5 6 7 8 9 1 2	son	☑
Horacio Homeowner	4 5 6 7 8 9 1 2 3	son	☑
			□
			□

If more than four dependents, see instructions and check here ▶ □

No. of children on 6c who:
• lived with you 2
• did not live with you due to divorce or separation (see instructions) 0

Dependents on 6c not entered above 0

d Total number of exemptions claimed . . . . Add numbers on lines above ▶ 2

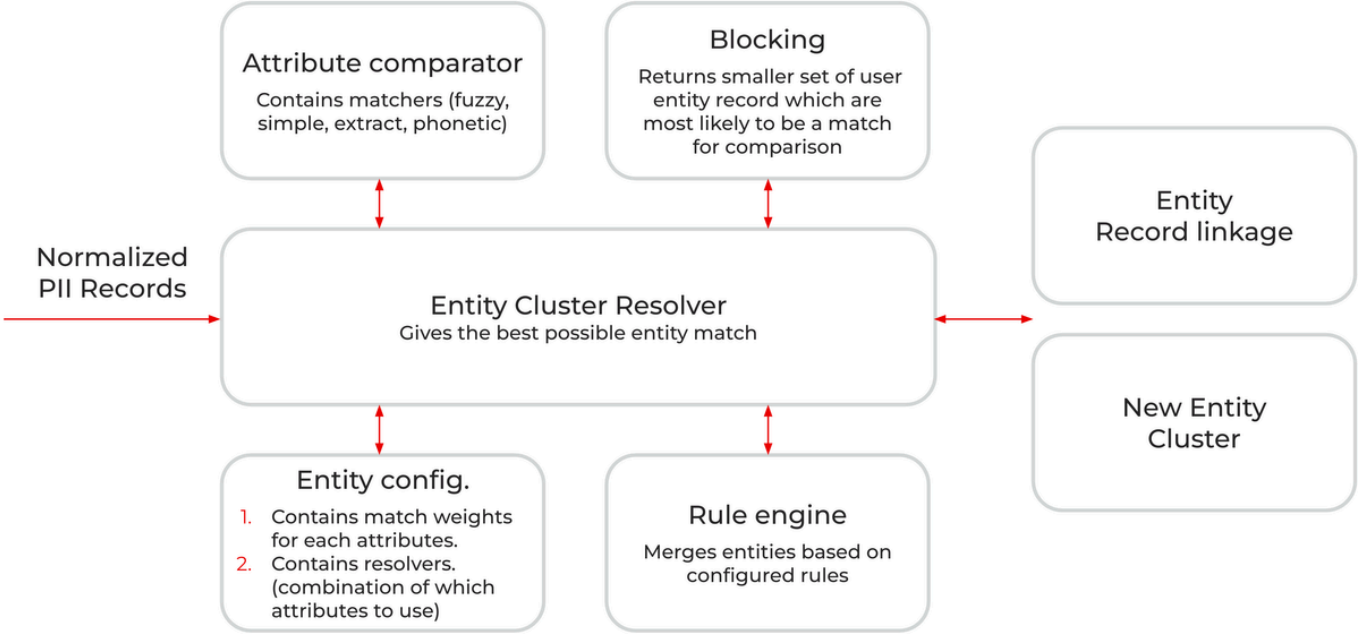
Income

Attach Form(s) W-2 here. Also attach Forms W-2G and 1099-R if tax was withheld.

If you did not get a W-2, see instructions.

7	Wages, salaries, tips, etc. Attach Form(s) W-2	7	98374
8a	**Taxable** interest. Attach Schedule B if required	8a	354
b	**Tax-exempt** interest. **Do not** include on line 8a . . . 8b		
9a	Ordinary dividends. Attach Schedule B if required	9a	
b	Qualified dividends 9b		
10	Taxable refunds, credits, or offsets of state and local income taxes	10	
11	Alimony received	11	
12	Business income or (loss). Attach Schedule C or C-EZ	12	12109
13	Capital gain or (loss). Attach Schedule D if required. If not required, check here ▶ □	13	
14	Other gains or (losses). Attach Form 4797	14	
15a	IRA distributions . 15a b Taxable amount . . .	15b	20117
16a	Pensions and annuities 16a b Taxable amount . . .	16b	16222
17	Rental real estate, royalties, partnerships, S corporations, trusts, etc. Attach Schedule E	17	-7,292
18	Farm income or (loss). Attach Schedule F	18	
19	Unemployment compensation	19	
20a	Social security benefits 20a b Taxable amount . . .	20b	

A US 1040 tax document with red bounding boxes showing how Lightbeam understands complex documents and the data inside of them.



Attribute comparator
Contains matchers (fuzzy, simple, extract, phonetic)

Blocking
Returns smaller set of user entity record which are most likely to be a match for comparison

Normalized PII Records

Entity Cluster Resolver
Gives the best possible entity match

Entity Record linkage

New Entity Cluster

Entity config.
1. Contains match weights for each attributes.
2. Contains resolvers. (combination of which attributes to use)

Rule engine
Merges entities based on configured rules

Entity Cluster Resolver workflow for identity resolution, normalized PII records flow through blocking and attribute comparison, guided by entity configuration and rule-based merging, to produce either entity record linkage or a new entity cluster.

To support rapid onboarding, Lightbeam is optimized for high-volume historical scanning. During initial deployment, the system schedules parallel processors to ingest existing content quickly and distributes resources dynamically across regions and cloud environments. At the same time, Lightbeam continuously monitors live changes and prioritizes new or modified content while the historical backfill runs in parallel, so teams get immediate visibility without waiting for the initial scan to complete.

This full-stack optimization from data ingestion to AI processing to entity resolution enables Lightbeam to operate efficiently at petabyte scale. Whether deployed in a global bank, healthcare network, or multinational enterprise, the system delivers continuous visibility, low-latency responsiveness, and high accuracy without requiring specialized hardware or proprietary infrastructure.

Scalability is not just about volume. It is about sustaining performance and precision at volume, while adapting continuously to change. That is what makes the Lightbeam Data Identity Graph different.

> ***"We chose Lightbeam because it offered us granular control and unique insights into our sensitive data— including the ability to figure out the identities associated with sensitive data—something no other solution matched."***
> ***- David Hanna, Director of Security, Veridian Credit Union***

The Connections That Power Governance

The real strength of the Data Identity Graph lies in its ability to uncover connections that reveal security risk(s) and enable actionable governance in real time. Its value is not just in identifying sensitive content, but in understanding who the data belongs to, who can access it, and whether that access is appropriate within the current business context.

Consider a real-world scenario involving a confidential merger and acquisition (M&A) project. A small group of executives, legal counsel, and finance leaders are working on the proposed acquisition of a publicly traded company. All documentation related to the deal is stored in a secure SharePoint folder with strict access controls. The project is highly sensitive and any unauthorized disclosure could lead to regulatory penalties, insider trading investigations, or massive fluctuations in stock price.

Despite best efforts, an analyst unknowingly puts a draft term sheet into a presentation slide and saves it in a team folder shared with the broader finance department. Traditional tools might scan the file and note the presence of financial figures or legal phrases but lack the context to recognize the severity of the exposure.

The Data Identity Graph detects that the file contains data tied to the confidential M&A project. It recognizes keywords, entity references, and relationships previously linked to the project's core documentation. It also identifies that the document has been stored in an unauthorized location and is accessible to users who are not part of the designated M&A working group.

Based on this analysis, Lightbeam triggers an automated governance response. Access is revoked for non-authorized users and incident response teams are alerted. A full audit trail is generated, including who created and accessed the file, when it was moved, and how it relates to the larger project entity. All of this occurs without requiring manual review or custom configuration.

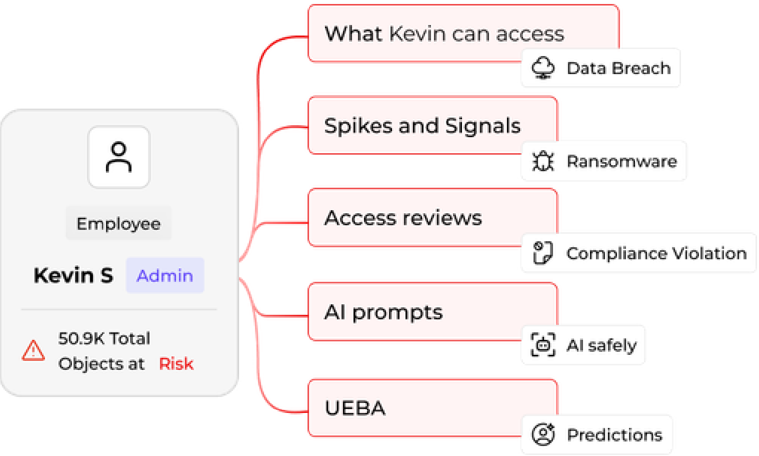
This kind of precision is only possible when sensitive content is mapped to data identities and business context in real time. By connecting the dots between data, people, and purpose, the Data Identity Graph transforms passive discovery into proactive risk mitigation and enables enforcement of business-critical governance policies.

Governance and Policy Enforcement

By mapping sensitive data to identity and access context, Lightbeam enables a new class of governance policies that go far beyond simple regex patterns or static data labels. These policies reflect real business logic and organizational rules, such as "only HR should access employee salary data" or "health insurance information that includes protected health information (PHI) must never be used in AI prompts or responses generated by AI agents."

These policies are enforced through automated playbooks that trigger based on the Data Identity Graph's understanding of the connections between people, data, and business purpose. For example, if a user attempts to upload a file containing employee health insurance records into a generative AI platform, Lightbeam can immediately block the action, revoke access to the file, alert security operations, and log the incident for audit and compliance purposes. Likewise, if an AI system attempts to respond to a prompt using sensitive PHI-related content, Lightbeam can intercept the response before it is delivered, preventing inadvertent exposure.

In another scenario, if a file containing expired customer data exceeds retention limits, the platform can trigger a predefined action such as redaction, archival, or deletion without requiring human review. The same policy-driven approach extends to consent management. If a customer has not granted marketing consent, Lightbeam can detect and flag marketing-related attributes and automatically enforce controls to prevent that data from being stored or retained across the environment.



A user profile and role context connect what the user can access, behavioral spikes and signals, access reviews, AI prompt activity, and UEBA insights to outcomes like data breach, ransomware, compliance violations, AI safety issues, and predictive risk.

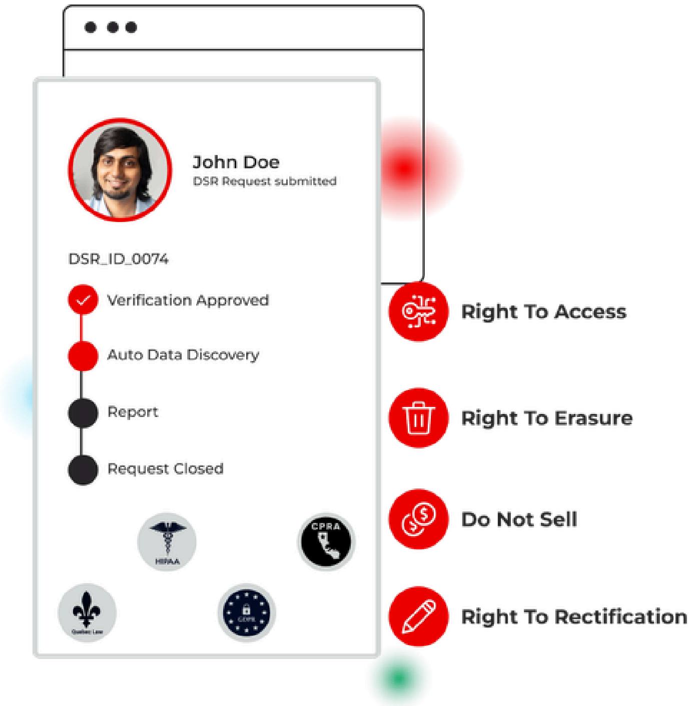
These governance policies are not built on static rules. They are informed by a continuously updated, identity-aware model of how data flows across people, systems, and processes. As users change roles, as new data is created, and as policies evolve, the Data Identity Graph adapts automatically to ensure that governance remains precise, dynamic, and aligned with business risk.

Enhancing Privacy Operations with the Data Identity Graph

Privacy operations are often among the most resource-intensive aspects of regulatory compliance. Responding to Data Subject Requests (DSRs) such as access, deletion, and correction under frameworks like GDPR, CCPA, and Quebec’s Law 25 requires organizations to locate every instance of an individual's data across a fragmented, constantly evolving data estate. For many enterprises, this remains a manual, error-prone process that increases compliance risk and erodes user trust.

The Data Identity Graph changes this reality by introducing entity-level precision and automation to privacy workflows.

A central challenge in privacy compliance is identity fragmentation. A single customer may be represented under different names, spellings, email addresses, phone numbers, or account IDs across sales, marketing, support, and product systems. Without a reliable method to unify these duplicate and disparate records, privacy teams are forced to guess, disclose too much, or overlook critical data, all of which create regulatory exposure.



A representation of the Automated DSR workflow in Lightbeam. A verified request triggers auto data discovery and reporting, then closes the case while supporting common privacy rights such as right to access, right to erasure, do not sell, and right to rectification across major regulations.

Lightbeam addresses this challenge through its patented entity resolution engine, which correlates identifiers such as emails, device IDs, phone numbers, names, and account numbers to construct continuously updated, unified profiles. When a DSR is received, the system instantly retrieves all relevant data tied to the requesting entity across structured, semi-structured, and unstructured systems with audit-ready accuracy. Completing a DSR manually, or even with a system that simply generates tickets and routes them to data owners for manual investigation, can cost as much as $1,500 per request. That cost is largely borne in employee time, lost productivity, and mounting backlog. It creates a hidden drain on privacy teams and business stakeholders. Lightbeam eliminates that cost entirely by automating the DSR workflow from end to end, including discovery, identity resolution, and execution of required actions such as data export or deletion.

Once the relevant data is located, automated actions such as redaction, deletion, or export can be executed according to predefined privacy playbooks. These workflows can be tailored based on sensitivity, jurisdiction, data type, or processing purpose, all without requiring manual review at each step. The result is a significant reduction in response time, often from weeks to minutes, while maintaining full traceability for compliance audits.

Consent management also becomes significantly more effective. Rather than relying on isolated cookies or device-level identifiers, Lightbeam anchors consent and legal processing bases to the unified entity. Whether a user withdraws consent through a website preference center or via a mobile app opt-out, Lightbeam ensures that the enforcement is consistent and immediate across all connected systems and departments.

The Data Identity Graph also enables automation of the Record of Processing Activities (ROPA), a core requirement of many global privacy regulations. Traditional ROPA generation involves surveys and spreadsheets that are static, incomplete, and outdated upon creation. Lightbeam transforms this process by generating a living, data-driven ROPA. As the system continuously scans the environment and links data to real entities, it dynamically classifies data subject categories such as employees, customers, or vendors. It can also infer the purpose of processing, such as payroll versus marketing, and help validate the lawful basis for that processing. This transforms the ROPA from a checklist exercise into an intelligent, continuously updated governance asset.

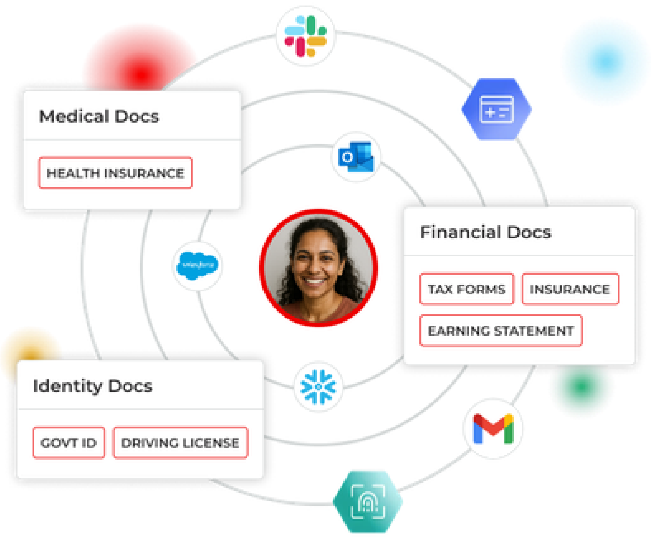
An additional benefit of the Lightbeam privacy operations capability is recurrence detection. If a data subject has been removed under a right-to-erasure request and later reappears in any system, intentionally or inadvertently, Lightbeam flags the reappearance for review. This safeguard ensures that privacy obligations are upheld over time, even in complex, distributed environments.

By eliminating identity ambiguity and automating discovery and enforcement, the Data Identity Graph delivers a scalable, future-proof foundation for privacy operations. It enables enterprises to meet evolving regulatory demands at petabyte scale with precision, speed, and minimal manual intervention.

Real-World Impact of the Data Identity Graph: Bank Data Security Transformed

The Data Identity Graph delivers operational value across a wide range of high-impact security and compliance use cases. A compelling example comes from a regional U.S. bank undergoing a SharePoint migration. Within 48 hours, Lightbeam ingested data from over 20 systems and resolved more than 189 million sensitive data fragments into mapped entities.

This immediate visibility enabled the bank to detect that sensitive mortgage documents had been inadvertently placed in a shared folder accessible to marketing interns, exposing regulated financial data. It also flagged that a retired employee continued to access account summaries weeks after departure, a clear access governance violation. In shared drives, it uncovered legacy folders containing protected health information (PHI) dating back to 2015, well past the organization's defined retention period. Each of these issues triggered automated remediation playbooks, including quarantining files, revoking unauthorized access, and generating audit logs.



Conceptual view of identity-centric data discovery. A single person is linked to sensitive document types such as identity, medical, and financial records across common SaaS and cloud repositories. Illustration only, not a product screenshot.

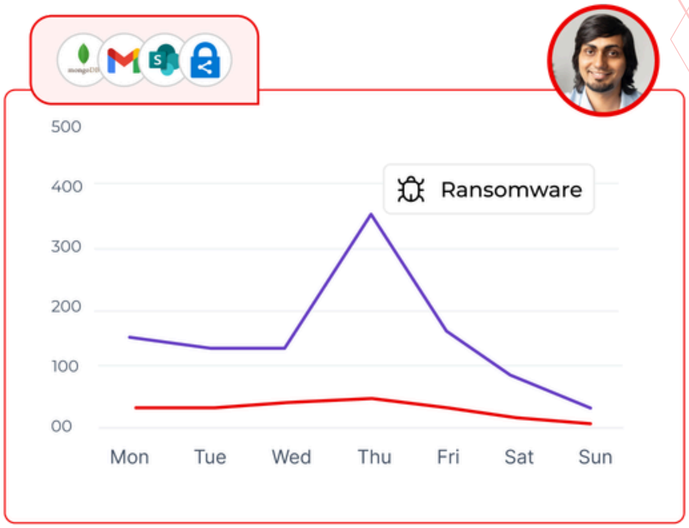
Beyond remediation, the bank used the Data Identity Graph to implement dynamic, identity-based policy enforcement. For instance, it configured access rules allowing support agents to view only the data related to open customer tickets assigned to them. This created a true least-privilege environment where access was continuously aligned to business needs.

In one case, Lightbeam identified a public S3 bucket exposing W-2 tax forms for 350 employees, including five C-level executives. Because the Data Identity Graph had linked these documents to specific individuals and job titles, the alert was automatically prioritized for immediate response.

Security teams quickly revoked public access, notified stakeholders, and documented the incident for compliance reporting.

For privacy compliance, the bank leveraged the graph to accelerate Data Subject Requests (DSRs). A right-to-erasure request for a former customer would trigger a graph query, locating and deleting all data tied to that person regardless of location, label, or format. The system also ensured that if new data tied to the erased subject reappeared in any system, it would be flagged for review to maintain ongoing compliance. This continuous monitoring ensured not only rapid resolution but also lasting adherence to privacy commitments.

In defending against ransomware, the bank used Lightbeam to detect and stop malicious encryption activity across all file types, whether or not they contained sensitive data. The system recognized behavioral anomalies in encryption patterns and acted to terminate the attack. For any files that were encrypted before containment, Lightbeam retained full knowledge of what sensitive information was inside, including the individuals or entities affected. This allowed the security team to assess whether breach notification thresholds were triggered under regulations such as HIPAA, GDPR, or state-level breach laws, and to respond appropriately with speed and clarity.



Illustrative ransomware detection and response. Lightbeam flags anomalous encryption behavior across repositories, helps contain the attack, and preserves context on what data and which identities were impacted to accelerate breach assessment and notification decisions.

The organization also used the Data Identity Graph to enforce defensible data retention policies. For example, data associated with customers whose last activity occurred more than seven years ago was flagged for deletion. This process resulted in the secure removal of over 80 terabytes of outdated customer information from cold storage, reducing both regulatory risk and infrastructure costs.

Insider threats were detected with high fidelity by correlating user behavior with sensitive data access. In one incident, a UK-based engineer accessed performance review files belonging to U.S. employees. Lightbeam flagged the incident immediately by linking the behavior to both abnormal access patterns and the sensitive nature of the files involved, triggering a security incident response.

These outcomes demonstrate how the Data Identity Graph does more than surface risks. It connects those risks to people, business context, and policy enforcement, empowering teams to take timely, targeted action. The result was not only improved visibility, but also measurable reductions in risk exposure, enhanced audit readiness, and operational cost savings across multiple business units.

How The Lightbeam Data Identity Graph Is Different

The Data Identity Graph was designed with a fundamentally different premise than most other tools in the data security and privacy space. While other platforms began as data catalogs, DLP tools, or access monitoring systems and later bolted on identity features, Lightbeam started from the opposite direction. We built our platform from the ground up around the concept of identity: who the data belongs to, how it connects to individuals or organizations, and how those relationships drive risk.

This architectural decision has far-reaching implications. Instead of treating data as the core asset and trying to analyze it for signs of sensitivity, Lightbeam treats identity as the anchor. Files, emails, messages, and records are simply extensions of the people they describe. This shift allows the system to answer questions others cannot. When a CISO wants to know, “Show me everything about John Smith,” the Data Identity Graph responds with a full picture, regardless of where the data lives or how it was labeled. Traditional systems, by contrast, are more likely to only be able to answer, “Where do we have SSNs?” and can’t connect that data to a real person.

Another key difference lies in how the system understands the context of access and risk. While some tools excel at showing who opened which files, Lightbeam goes further by also identifying whose data was accessed. For example, if a contractor views a spreadsheet, the platform will not just log the user and the file. It will reveal that the document contained the CEO’s health records or a customer’s financial account. This context changes how incidents are triaged, how alerts are prioritized, and how responses are triggered.

Precision is built into every layer of the identity resolution process. The system does not rely on basic string matches or static rules. Instead, it assigns a dynamic confidence score to every potential match, weighing rare identifiers more heavily than common ones. For example, a match on an employee ID or government-issued number carries more weight than a match on a first name. Multiple data points, such as name, email, and phone number, compound confidence when they all converge on the same person. And if any attribute connects to a verified source of truth, such as a system of record, the confidence level increases dramatically.

This fine-grained scoring ensures that Lightbeam only forms identity connections when there is sufficient evidence. If the confidence level falls below the threshold, the system does not guess. Instead, it creates a provisional, unverified profile that remains available for future matching as additional evidence becomes available. In cases where the data matches overly common patterns—such as shared support email addresses or generic job titles—it is excluded from identity graph construction to avoid false positives and maintain the integrity of the graph.

However, this data is still retained within Lightbeam's platform. It remains indexed and available to support future recall, enabling the system to revisit prior findings when more context or identifiers become available. This approach allows the graph to evolve dynamically and improve over time without introducing misleading or irrelevant relationships.

Lightbeam also handles ambiguity carefully. While some platforms attempt to merge possible duplicates automatically, Lightbeam prioritizes accuracy over assumption. Users have full control over merging duplicate profiles and are given tools to manually review and consolidate data identities when needed. Once confirmed, all data linked to either profile is reassigned to a unified entity, enabling more comprehensive and accurate governance.

This commitment to clarity over completeness is intentional. Rather than build an identity graph that is fast but noisy, Lightbeam has engineered one that is deliberate, trustworthy, and highly scalable. This focus enables precision in downstream use cases, from automating access controls to enforcing privacy rights and responding to security events.

In short, Lightbeam does not simply label data, it understands it. It does not just show who touched what, it reveals whose data was touched and why that matters. And it does not try to approximate identity, it builds it from the ground up with the rigor and context that modern data security demands.

Future Direction and Innovation

Lightbeam is evolving the Data Identity Graph from a descriptive, real-time mapping layer into an autonomous, predictive engine that actively governs and protects sensitive data across complex, AI-assisted environments.

The next generation of the Data Identity Graph will introduce several breakthrough capabilities:

End-to-End Data Lineage Analysis

The Lightbeam Data Identity Graph will not only understand where data resides and who it belongs to, but also how it flows through the organization. This includes tracking how sensitive data is created, copied, transformed, and shared, automatically establishing relationships between source and derivative content.

For example, if a financial analyst copies a table from a spreadsheet named *Q3_Financials.xlsx* and pastes it into a PowerPoint presentation, the graph will create a "copied_from" link between those two objects. If that presentation is later emailed to a C-level executive, the graph can reconstruct the complete path of the data: “The sensitive customer data in this presentation originated from *Q3_Financials.xlsx*, created by analyst_A, and was emailed to executive_B at 4:30 PM.” This type of lineage mapping is invaluable for forensic investigations, breach containment, and understanding the lifecycle of high-risk information.

As organizations increasingly deploy agentic AI systems that generate, summarize, and store content automatically, tracing lineage becomes even more critical. Sensitive data can inadvertently propagate through AI training processes, memory states, or generated outputs. By capturing the full journey of data across both human and machine interactions, the Data Identity Graph will provide the accountability and traceability needed to govern AI systems and prevent unintentional data leakage.

Autonomous Governance and Remediation with LLMs

The next generation of the graph will also function as the decision-making core of an automated data security platform. Rather than waiting for human review or static rules, the graph will autonomously detect risk and trigger remediation in real time.

One key advancement is the integration of Large Language Models (LLMs) to detect “policy drift”, the slow, natural leakage of sensitive content from secure locations into general-purpose collaboration tools.

Consider a confidential initiative like Project Phoenix, originally protected in a secure SharePoint and private Slack channel. Over time, project content may be copied into emails, pasted into public chats, or replicated in docs with broader access. Traditional systems fail to detect this because they rely on manually applied labels or hard-coded rules.

The LLM-enhanced Data Identity Graph learns the “DNA” of a project—its language, concepts, and artifacts, from its secure environment. It then monitors other systems for signs of related content appearing in unexpected or unauthorized locations. If such drift is detected, the system can evaluate the cumulative risk and trigger alerts, escalation, or access restrictions automatically.

Trusted Breach Attestation Protocol

Finally, Lightbeam envisions a new inter-organizational protocol for breach notifications. When both parties in a data-sharing relationship use the Data Identity Graph, a breach can be communicated instantly and privately, without ever exposing the raw list of affected individuals.

Instead of sending sensitive spreadsheets of breached users, the breached system sends a cryptographic attestation that mathematically proves which of your customers or employees were impacted. Your system validates the proof and takes immediate, automated action to notify internal teams, revoke access, or launch investigations. It transforms breach response from a legal negotiation into a trust-based, automated protocol that strengthens partnerships.

Looking Ahead

As business environments become more hybrid, multi-cloud, and AI-driven, the Data Identity Graph is expanding its role. It will serve as the connective tissue between identity, data, and policy—providing not just awareness, but intelligent enforcement and continuous adaptation.

The Lightbeam vision is clear: to deliver an always-accurate, identity-centric foundation for protecting sensitive data in motion, in use, and at rest, regardless of where it lives or how it evolves.

Conclusion: Data Identity Is the New Perimeter

In the post-perimeter era, where data moves freely and AI drives decision-making, protecting sensitive information requires more than surveillance. It requires understanding. The Lightbeam Data Identity Graph brings that understanding to life, connecting the dots between people, data, and access in ways that empower automated governance at scale.

Where legacy tools generate alerts, Lightbeam generates action. Where others look at files, we look at people. And where others offer policies, we deliver precise enforcement.

For organizations serious about data security in the AI age, the Data Identity Graph is not just an upgrade. It's a new operating model.

Glossary of Terms

Attribute

A data value, portion of a data element, or field associated with an entity or data identity, such as name, email, phone number, or account ID. Attributes are used by Lightbeam to identify, classify, and associate data with specific entities and data identities. Also referred to as data elements.

Confidence Score

A numeric value generated by the Lightbeam Data Identity Graph that indicates match likelihood and classification certainty. Confidence scores apply to entity resolution, to specific high-sensitivity attributes, and to document classification. Higher scores reflect stronger signals based on unique identifiers, corroborating attributes, and verified sources.

Control Plane

The component of the architecture responsible for centralized configuration, reporting, and orchestration. It coordinates scanning operations across environments.

Data Element

An individual, identifiable piece of information, such as a Social Security number, medical record number, or device ID. Data elements are parsed, classified, and associated with entities to support governance and privacy use cases. Often used interchangeably with attributes.

Data Identity Graph

The core intelligence layer of the Lightbeam platform. It dynamically links fragmented data elements to entities (data subjects or accessors), enabling identity-aware visibility, classification, access control, and policy enforcement across structured, semi-structured, and unstructured environments.

Data Identities

The representations of people, departments, companies, or digital assets to which sensitive information belongs. Data identities are created through the correlation of attributes and context, enabling Lightbeam to apply governance at the entity level.

Data Plane

The distributed component of the system deployed close to the data source. It performs local scanning, parsing, and classification without moving data across environments. Data planes can be independently deployed across clouds or on-premises infrastructure.

Data Subject

The person, group, or object described by the data. A data subject may be a human (e.g., employee, patient) or non-human (e.g., department, company, contract, or file). Data subjects are represented in the system as entities.

Data Subject Request (DSR)

A formal request from an individual under laws like GDPR or CCPA to access, delete, or correct their personal data. Lightbeam automates DSR fulfillment by linking all relevant data to unified entities and applying predefined privacy workflows.

Entity

A foundational concept in the Data Identity Graph representing any data subject (human or non-human) or accessor. Entities can include people, companies, files, contracts, departments, and more. They are constructed by correlating attributes and enriched with context to support governance, privacy, and security.

Entity Resolution

The process by which Lightbeam unifies disparate or fragmented records into a single entity by matching attributes with high confidence. It accounts for variations, duplicates, and ambiguous data across systems.

Large Language Model

A type of advanced AI model used selectively for deeper contextual understanding. LLMs are deployed locally to analyze complex documents and relationships without exposing sensitive data outside the network.

Retrieval-Augmented Generation (RAG)

An AI method used to enhance classification by retrieving stored document templates and applying them to similar content. RAG supports precise field-level extraction and contextual understanding in structured and unstructured data.

Record of Processing Activities (RoPA)

A regulatory requirement under laws like GDPR to maintain documentation of how personal data is processed. Lightbeam automates ROPA creation using real-time entity relationships, inferred processing purposes, and dynamic classification.

System of Truth (SOT)

An authoritative source of identity and business data, such as HR systems, CRMs, or identity governance platforms. Matches against SOTs carry greater weight in the confidence scoring model.

Unstructured Data

Data that lacks a fixed schema or format, such as emails, chat messages, documents, and PDFs. The platform parses unstructured content to extract data elements and relationships that inform governance decisions.

Lightbeam

Turn identity into your strongest data control.

Ready to turn insight into action? Get a free demo of the Lightbeam Data Identity Graph and start enforcing data policy with confidence.

Get Demo

About Lightbeam

Lightbeam is an identity-centric data security platform that reduces breach risks, ransomware costs, and regulatory penalties by unifying data security posture management (DSPM), privacy, and governance. Powered by patented Data Identity Graph technology, Lightbeam discovers and maps sensitive data across structured, unstructured, and semi-structured sources, including shadow data, and links it to human identities and business context. This visibility helps organizations enforce data access policies, automate privacy workflows such as DSR, RoPA, and consent, and reduce risk through precise redaction, archival, deletion, and access governance. Lightbeam enables customers to simplify data security and protect privacy without slowing down the business.

@Lightbeam 2026

https://lightbeam.ai